Functional genomics in *Arabidopsis*: large-scale insertional mutagenesis complements the genome sequencing project

Serguei Parinov and Venkatesan Sundaresan*

The ultimate goal of genome research on the model flowering plant *Arabidopsis thaliana* is the identification of all of the genes and understanding their functions. A major step towards this goal, the genome sequencing project, is nearing completion; however, functional studies of newly discovered genes have not yet kept up to this pace. Recent progress in large-scale insertional mutagenesis opens new possibilities for functional genomics in *Arabidopsis*. The number of T-DNA and transposon insertion lines from different laboratories will soon represent insertions into most *Arabidopsis* genes. Vast resources of gene knockouts are becoming available that can be subjected to different types of reverse genetics screens to deduce the functions of the sequenced genes.

Addresses
Institute of Molecular Agrobiology, 1 Research Link, The National University of Singapore, Singapore 117604
*e-mail: director@ima.org.sg

Current Opinion in Biotechnology 2000, 11:157–161

0958-1669/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved.

Abbreviations
Ac/Ds  Activator/Dissociation
En/Spm  Enhancer/Suppressor-mutator
I  Inhibitor

Introduction

*Arabidopsis thaliana* has been the subject of intense research for over a decade as a model organism for the molecular genetics of flowering plants because of its short generation time, small genome and high gene density. A very important step of this research is the *Arabidopsis* Genome Initiative (AGI) to determine the genome sequence of *Arabidopsis*. Currently, over 80% of the predicted 130 Mb of *Arabidopsis* genome has been sequenced, including complete sequences for two out of five *Arabidopsis* chromosomes, 2 and 4 [1,2]. It is expected that the entire genome sequencing will be completed before the end of the year 2000, thus making this the first genome of a flowering plant to be sequenced.

The utility of the genome-sequencing project relies significantly on the prediction of coding regions in the genome. Much of the genomic sequences of bacterial artificial chromosome (BAC) and P1 clones submitted to GenBank are annotated with the results of gene prediction, produced by a combination of different computer algorithms. Such annotation is often a short summary and it does not always unambiguously reflect the real gene sequences. Nevertheless, half of the predicted genes match expressed sequence tag (EST) and cDNA sequences [3]. Alternative sources of gene annotation are also being developed. For example, the *Arabidopsis* Genome Displayer (KAOS; http://www.kazusa.or.jp/kaos) and the *Arabidopsis* Genome Annotation Database at the Institute for Genomic Research (http://www.tigr.org/tdb/ath1/htmls/ath1/html) present recent updates of the annotation for sequenced *Arabidopsis* clones. The convenient graphical outputs provided by these sites help users to make independent decisions about the gene structure in a selected region of genome by comparison of the predictions made with different gene prediction algorithms.

According to the analysis of sequenced genomic regions on chromosomes 2, 4 and 5 [1–3], ~50% of the *Arabidopsis* genome encodes genes, with an average density of about one gene per 4.5 kb. Therefore, the complete ~130 Mb *Arabidopsis* genome should carry about 25,000–30,000 genes, most of which have not been studied yet. The completion of the genome project, therefore, opens a new frontier for functional genomics. The accumulation of sequence and functional information about *Arabidopsis* genes will promote the application of reverse genetics methodologies to decipher the functions of novel genes that are homologous to known genes, or to study unknown genes that interact biochemically (e.g. in a yeast two-hybrid assay) with a gene of known function. This review covers recent progress in the application of large-scale insertional mutagenesis strategies for functional genomics in *Arabidopsis*.

**Gene inactivation: site-specific versus random insertional mutagenesis**

The complete inactivation of a gene is generally the most direct way to understand its function. Unfortunately, there are no effective methods for targeted gene replacement in flowering plants. Nevertheless, development of novel tools for site-specific mutagenesis continues. Beetham *et al.* [4], for example, successfully used self-complementary chimeric oligonucleotides to introduce site-specific base substitution in a nuclear gene of *Nicotiana tabacum*. This approach, however, has to be improved significantly in terms of efficiency for use in functional genomics. Approaches such as antisense and co-suppression have been used effectively to inactivate genes of known sequence [5]. These strategies are also not efficient enough to be used on a large scale, and presently are mostly limited to the study of single genes.

Currently, the most effective alternative strategy is random large-scale insertional mutagenesis. The high gene density [3] in the *Arabidopsis* genome is particularly favorable for random mutagenesis, as every second insertion will disrupt
a gene sequence. The insertion of large T-DNA or transposon constructs that are several kb in length most often leads to complete inactivation of a gene. Because the sequence of the insertion is known, primers complementary to the end of insertion and to the gene of interest can be used for PCR screening of insertion collections to identify the plants carrying an insertion within the gene of interest. Alternatively, the disrupted gene can be identified directly by sequencing of the DNA flanking each insertion [6]. The completion of the genome-sequencing project will enable any short flanking sequence to be used to find the exact position of the insert within sequenced genomic DNA, which then can be analyzed to identify the disrupted gene.

**Systems for insertional mutagenesis**

**T-DNA as a mutagen**

Agrobacterium-mediated T-DNA transformation has been used for insertional mutagenesis in different plant species. Plant transformation by Agrobacterium results in the integration into the nuclear genome of a sequence called T-DNA, which is carried on a bacterial plasmid. As the use of T-DNA as a mutagen in Arabidopsis has recently been reviewed in detail elsewhere [7••], only a brief summary is presented here. Large collections of T-DNA transformants of Arabidopsis have been collected independently by several groups and have been used extensively for reverse genetics [7••,8,9]. One advantage of using T-DNA as an insertional mutagen is that it directly generates stable insertions into genomic DNA and does not require additional steps to stabilize the insert. A second advantage is that T-DNA insertion appears to be completely random, as no T-DNA integration hot spots or integration preferences have been reported. On the other hand, complex patterns of T-DNA integration, including transfer of vector sequences adjacent to T-DNA borders and the large frequency of concatemeric T-DNA insertions, can complicate further PCR analysis for reverse genetics [10,11]. Small and major chromosomal rearrangements induced by T-DNA integration have been frequently observed, leading to difficulties in the genetic analysis of the insertion, such as mutant phenotypes that are not correlated with the T-DNA insertion [12,13]. Nevertheless, the development of highly efficient methods of T-DNA transformation for Arabidopsis [7••,14] has made it feasible for laboratories to generate thousands of transformants relatively rapidly, and as a result T-DNA is now a widely used insertional mutagen in Arabidopsis.

**Transposon mutagenesis**

Two systems of heterologous transposable elements, both originating in maize, have been successfully used for insertional mutagenesis in Arabidopsis: Activator/Dissociation (Ac/Ds) and Enhancer/Suppressor-mutator (En/Spm). There are two advantages of using the transposons for insertional mutagenesis. First, unlike T-DNA, a transposon can be excised from the disrupted gene in the presence of transposase, with the resulting reversion of the mutation. This provides a simple means to confirm that an observed mutation is tagged by the transposon [15,16••,17••].

Second, most transposition events occur preferentially to linked sites [18,19]. A transposon insertion library generated from any donor transposon will therefore be enriched for insertions into the adjacent genes. If there is a transposable element near to the gene of interest, it can be efficiently mobilized to reinsert within the gene. This strategy has been successfully applied by Ito et al. [19] and Seki et al. [20] for regional Ac/Ds insertional mutagenesis of the genes from CIC7E11/8B11 and 5CIC5F11/CIC2B9 loci on Chromosome 5. These authors focused their attention on the cDNAs specific to these chromosomal regions, which were selected using a cDNA-scanning method that allows one to selectively clone cDNAs from any small region of genome. PCR analysis of DNA flanking the insertions showed that 14–20% of transpositions were actually located within the target regions, each representing about 1 Mb of genomic DNA surrounding the Ds donor sites.

On the other hand, preferential short-range transposition is an obvious drawback if the target goal is to achieve random saturation of genome with insertions. To address this problem, a two-element Ac/Ds system was developed in which negative selection against closely linked transpositions was employed [21••]. Simultaneous negative selection against the Ac transposase permitted selection for plants that carry stable Ds insertions. Using this system, two major Ds transposition hot spots located in the narrow regions adjacent to nucleolar organizer regions on chromosome 2 and 4 were found, indicating that these regions of the genome are preferential targets for Ds insertions. It has been demonstrated that Ds elements transpose preferentially to 5′ regions of genes, therefore increasing the chances of obtaining complete loss-of-function mutations, which are simpler to interpret.

A similar system that utilizes En/Spm transposable elements was designed by Tissier et al. [22••] in order to simplify the process of generating transposants. In this system, selectable markers that allow selection of soil-grown plants were employed to reduce the labor associated with growing plants on media and transplanting. Another modification is that in contrast to the Ac/Ds system described above in which the initial transformants carrying Ac and Ds are maintained as separate lines [21••], the mobile dSpm is combined together with the Spm transposase in the same T-DNA so that negative selection has to be applied against only a single locus, thereby reducing the number of progeny to be screened by a factor of four. In this system, it is not possible to maintain the primary transformants as defined starter lines, because the dSpm elements will be continually transposing in the presence of Spm transposase. Instead of maintaining and using several starter lines, a large number of independent primary T-DNA transformants were generated as sources of transpositions. Availability of simple T-DNA transformation methods...
makes this strategy feasible in *Arabidopsis*, though not in most other plants.

The two systems described above were designed to select predominantly stable single insertions. Hence further mutant analysis of these lines is relatively straightforward. Also, each individual insertion line can be rapidly characterized by direct sequencing and analysis of DNA flanking the insertion [21••,22••]. Notwithstanding the above advantages, it is necessary to generate more than 100,000 individual insertion lines for saturation of the genome.

An alternative approach for insertional mutagenesis using plants carrying multiple insertions of a mobile *Inhibitor I* (dSpm) element has been used, which permits genome saturation using a relatively low number of lines [16••]. Seven ‘founder’ lines were used as starters each carrying 7–10 independent inserts of *I* (dSpm) and a source of *En* (Spm) transposase that is required for transposition of the *I* (dSpm) element. After several generations, a ‘multiple transposon population’ of 2,592 lines, each carrying 20–25 *I* elements, was generated, corresponding to 50,000–65,000 insertions in total. Because of the high gene density in *Arabidopsis*, 20 insertions in a plant will probably result in inactivation of multiple genes, which could result in complex mutant phenotypes. Nevertheless, due to the extensive functional redundancy in the *Arabidopsis* genome [7••,17••,23], this method can be effectively utilized for studies of genes that produce unambiguous phenotypes when Knocked out, as demonstrated by Speulman et al. [16••]. The presence of the *En* transposase in these lines allows the identification of revertants within the same generation. Characterization of lines that carry multiple insertions of mobile transposons frequently requires many extra steps compared to single-copy insertion lines, especially when the phenotype of a gene knockout is not clear. For stabilization of the insertion, as well as for reduction of transposon copy number, these lines would have to be subjected to subsequent crosses to wild-type plants. The genetic analysis of the observed mutants is also complicated by the presence of multiple insertions and transposon footprints that result in gene inactivation with no associated transposon. In addition, a background of somatic insertions complicates PCR screening strategies to identify gene knockouts, as some of these insertions can generate the correct PCR products, but will not be germinally transmitted. Some of these drawbacks could be overcome by generating a much larger collection of multiple transposon insertion lines, which would increase the probability of having several independent insertions into the same gene, which can be used for confirmation of the mutant phenotypes.

**Screening strategies for reverse genetics**

For effective reverse genetics, it is preferable to have as large a collection of insertions as possible. Assuming a 130 Mb genome size, it is estimated that at least 120,000 random insertions are required to tag any 5 kb *Arabidopsis* gene with 99% probability. The screening of such collections in order to obtain an insertion in a particular gene of interest is usually performed using PCR-based approaches. Two basic strategies have been applied for the screening of insertion lines.

The first so-called ‘pooling strategy’ is to combine 20–100 insertion lines together to form a pool. DNA extracted from these pools is then used to perform the PCR screening using gene- and insert-specific primers [8,17••,22••]. It reduces the work associated with DNA preparation and PCR screening by a factor equal to number of lines in a pool. Individual lines from the PCR-positive pool can be subsequently examined to identify the line carrying the required insertion. As the number of pools required for genome saturation is also high (for example 1000 pools of 100 lines each for a collection of 100,000 insertions), DNA from pools can also be combined into superpools and the PCR screening can be performed in stages [7••,22••]. Alternatively, plants can be pooled according to a three-dimensional gridding strategy that further reduces the number of DNA pools for the PCR screen [16••]. As an example, a collection of 1000 insertion lines can be pooled into a grid consisting of 10 rows, 10 columns and 10 blocks, so that an individual line is identified by an address of one row, one column and one block. Thus only 30 DNA samples have to be prepared and subjected to PCR to identify the exact line carrying the desired insertion within the population of 1000. A modified format of the pooling strategy called ‘inverse display of insertions’, designed to compact and simplify distribution of the DNA libraries, has also been tested [17••,22••]. In this method, the multiple fragments flanking the insertions from the DNA pools are amplified by PCR, and the products immobilized on membranes. The screening is then performed by hybridization of the DNA of interest with the membranes carrying the array of DNA samples. The advantage of pooling strategies is that they are not very labor intensive and they provide a fast short-term resource for researchers planning to perform their own PCR screening of available collections.

An alternative strategy is for sequences flanking every individual insert to be amplified and sequenced and placed in a catalogue [21••,22••]. In this case, screening for a mutant line is simplified to a search for a corresponding insertion site in the catalogue. Obviously, this strategy can only be used for insertions that can be stably propagated. The large-scale application of this strategy requires much more time and effort than the pooling strategy. Random sequencing of every insertion line can be convenient and economically attractive over the long run, however, because it provides precise information and eliminates the subsequent labor associated with DNA distribution and PCR screening. The detailed analysis of every insertion site can be omitted if all flanking sequences are organized into a basic local alignment search tool (BLAST) dataset, so that the actual screening operation is a BLAST search of such a dataset for a sequence identical to the sequence of a given gene.
Current availability of insertions covers at least half of all Arabidopsis genes

Large collections of T-DNA and transposon insertion lines have recently been produced independently by different laboratories. An important point is that the number of DNA samples extracted from these lines is also substantial. For example, DNA pools of 48 from 35,000 T-DNA insertion lines (55,000 inserts) [8], 960 DNA pools (also combined into large superpools) representing 48,000 independent stable dSpm transposants [22**], and pools of DNA from 2592 multiple insertion lines (representing 50,000 to 65,000 insertions of fl(dSpm)) [16**] have been generated. Many of these lines have been sent to The Arabidopsis Biological Resource Center (ABRC) and Nottingham Arabidopsis Stock Center (NASC). Additionally, an Arabidopsis knockout facility has recently been established at the University of Wisconsin to provide access for the PCR screening of 60,480 T-DNA insertion lines [7**]. In parallel, identification of the stable insertion positions by sequencing and analysis of DNA flanking the inserts has also progressed [16**,20,21**,22**]. For example, a database of 1200 dSpm Sequenced Insertions Sites (SINS) from pools of transposants [22**], and a database of 500 individual Ds insertion lines with identified positions [21**] have been established. Both databases are expected to expand significantly in the near future.

A practical evaluation of some of the available collections of DNA samples for reverse genetics screening has been performed by Meissner et al. [17**]. They used PCR- and hybridization-based methods to identify insertions in Arabidopsis genes of the MYB family of transcription factors. The collection of insertion lines they used for screening included 13,264 T-DNA lines, 28,800 stable dSpm insertion lines, 8000 lines containing ~50,000 autonomous En/Spm insertions, and 2592 En/dSpm lines containing ~20–25 mobile dSpm insertions. In total, the screened lines represent roughly 150,000 insertion events that are estimated to hit 85% of the ~1.6 kb long MYB genes, assuming a random distribution of inserts. In fact, 47 insertions into 36 genes out of a total of 73 MYB members (50%) were isolated. Assuming the median size of an Arabidopsis gene is 2.1 kb [7**] these results indicate that effective saturation of more than 50% of genes was achieved in the experimental population.

Conclusions

The intensive efforts in the past few years to undertake insertional mutagenesis in Arabidopsis are beginning to make a major impact on the functional genomics of this plant. Large numbers of T-DNA and transposon insertion lines have been generated by different groups, and if combined together these collections may already represent insertions into most Arabidopsis genes [7**,8,9,16**,21**,22**,24,25]. The Arabidopsis Biological Resource Center (ABRC) of the Ohio State University (http://aims.cps.msu.edu/aims) [24] in cooperation with Nottingham Arabidopsis Stock Center (NASC; http://nasc.nott.ac.uk) [25] are organizing the collection, preservation and distribution of transgenic seeds acquired from these sources. DNA sample collection from some of these lines has also been initiated in order to enable screening for insertions in particular genes of interest by PCR [7**,8,16**,22**] or hybridization [22**]. As an alternative to PCR screening, databases of insertion sites have been established by laboratories that are sequencing DNA flanking the insertions [21**,22**].

Because of the high gene density, insertional mutagenesis is a powerful tool in Arabidopsis — every second insertion can be found within an Arabidopsis gene. Despite growing collections of insertions, saturation insertional mutagenesis may remain an elusive goal, especially for small target-gene sizes. For example, to find an insertion in a 1 kb gene with 99% probability, 600,000 insertion lines would need to be generated and screened [7**].

In cases where the desired insertion is not found by screening the available collections of insertions, a gene can still be tagged effectively using a closely linked transposon insertion [19,20,21**]. The donor line containing the closely linked transposable element can be selected from the collections of mapped transposons using the sequence databases [21**,22**] and crossed with plants expressing the transposase to remobilize the element.

The above resources build a strong foundation for further functional characterization of inactivated genes. Many Arabidopsis genes demonstrate different levels of functional redundancy, whereas others are only needed for survival in specific conditions, such as resistance to specific biological or environmental stresses [7**]. As a result, most of the insertion lines do not demonstrate clear morphological phenotypes when grown and examined under standard conditions [7**,8,17**,23]. The redundancy problem can be addressed by constructing double mutants with closely related genes. Identification of knockouts in such genes by either the pooling strategy or the flanking sequence database strategy will be an important step in this approach. In the case of conditional phenotypes, examination of mutant plants under a variety of experimental conditions will be required to decipher gene function [7**]. In this case, homology of the disrupted gene to known genes can help to predict the proper experimental conditions. In either case, functional studies of the genes will be greatly facilitated by knowledge of their expression patterns and regulation. The use of gene/ enhancer trap elements can provide this information for individual genes [23,26]. On a larger scale, this information will be considerably enhanced by the use of new technologies such as microarrays (see Schaffer et al., this issue pp 162–167).

Acknowledgements

We thank the National Science and Technology Board of Singapore for supporting our research.
Functional genomics in Arabidopsis Parinov and Sundaresan 161

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
• of special interest
•• of outstanding interest


A different application of the Enhancer-Inhibitor En1 (also known as Suppressor-mutator Spm-dSpm) transposon system for functional genomics is described. Seven original ‘founder’ (started) lines carrying 7–10 mobile (dSpm) elements and an En transposase gene were used to generate a series of mutant lines by single-locus seed such that multiple (20–25) insertions occurred per plant. ‘Multiple transposant population’ of 2662 lines were produced and combined in pools. DNA from the pools was extracted and can be used for PCR screening using a three-directional strategy.


The authors present results of molecular analysis using a collection of lines containing independent Dissociation (D) element insertions distributed throughout the Arabidopsis genome. Distribution of Ds transpositions showed significant (~10%) preference to the regions adjacent to nuclear organizer regions on chromosomes 2 and 4, but were otherwise random. The sequences of over 500 lines with characterized insertion position by reference to GenBank sequences are presented as a database, with the corresponding lines available from Nottingham Arabidopsis Stock Centre http://nasc.nott.ac.uk/ima.html.


